Discussion for a statement for biobank and cohort studies in human genome epidemiology

John P.A. Ioannidis, MD

International Biobank and Cohort Studies meeting

Atlanta Feb 7-8, 2005

Key elements in prior relevant statements

- Participant flow
- Recruitment
- Baseline data
- Numbers analyzed
- Outcomes and estimation
- Ancillary analyses

Key issues to consider: 1

- Results should have a perfect symmetry with the Methods. What is promised in the Methods should be shown in the Results and what is shown in the Results should be described in the Methods

Key issues to consider: 2

- Reporting should be comprehensive; selective reporting (often downgraded to “best results reporting” but not necessarily) should be avoided, as it leaves tremendous potential for bias (=“report exactly what you did and everything that you did”)

Key issues to consider: 3

- Given that cohort studies may have a lot more ancillary analyses than randomized trials, and genetic epidemiological studies may have more ancillary analyses than many other epidemiological studies, accompanying web supplements of raw data or web supplements of detailed analyses should be considered

Key issues to consider: 4

- Reporting should help understand better the potential extent of selection bias, confounding, and information bias (misclassification). A good study should be very transparent about these potential biases and should highlight them rather than hide them in a vain effort to look “perfect”

Participant flow

- Very essential
- Flow diagram would be useful with stages of subjects invited/sought – recruited – selected for analysis (including how, e.g. sampling methods) – followed-up in the cohort (with quantitative information on extent of censoring e.g. at a minimum a summary measure like available follow-up divided by total possible follow-up in the absence of any censoring, or more detailed +/-graphical information) – analyzed with descriptions of deviations/losses at each stage
- Provide explanation of reasons for these deviations/losses at each stage

Recruitment

- Dates for periods of recruitment and follow-up may be less important than for clinical trials, but may still be worthwhile reporting
- Important to document the recruitment process from the point of what selection biases may have operated to form first the recruited population and then the analyzed population, and whether these selection forces are expected to potentially distort the results (e.g. geographical area, racial-ancestry restrictions, any sampling methods used, recruitment rates over time)
- When recruitment has been different for subgroups/subpopulations, each subgroup/subpopulation should be described separately

Baseline data

- Baseline demographic and clinical characteristics, as well as characteristics that are considered potentially important for modifying genetic effect sizes (consider in particular gene-environment interactions) overall and per genotype or haplotype being analyzed
- Hardy-Weinberg check on all genotypes
- Data on ancestry (and mode of definition/ascertainment), spectrum of disease severity (with clear information on definitions/ascertainment), characteristics that are helpful to evaluate/probe the potential misclassification rate of diseased and healthy subjects, any relevant family history and familial relatedness of subjects (if pertinent)

Numbers analyzed

- Numbers of participants (denominators) in each group and in each analysis
- Absolute numbers should be given whenever feasible (consider also web supplement)
- While cohort studies typically use relative risks (specify whether hazard ratio from Cox model, relative rate from Poisson model, risk ratio from crude numbers, incidence rate ratio from crude numbers with person-years, other) some absolute measures of effect should also be considered for presentation. These could include absolute event rates at different time points, with full information shown with K-M curves with data per genotype/haplotype of interest

Outcomes and estimation

- For each primary and secondary analysis, summary of results of each group and effect size along with precision estimate (95% CI)
- Useful to have the 2*2 or 2*k tables presented also for the unadjusted analyses (plus the adjusted or interaction analyses, if these are defined as the primary analyses). For cohorts, denominators may have to be person-years rather than persons. The essential points are to (1) allow replication of the main calculations and (2) allow meaningful use of the data in future data synthesis

Ancillary analyses

- Address multiplicity (major issue in epidemiological studies)
- Describe all subgroup analyses undertaken
- Describe all adjusted analyses undertaken with details on how adjustment was performed
- Describe all interaction analyses undertaken with details on how interactions were defined and selected
- Describe all sensitivity analyses undertaken to examine bias or other problems with the data (common in epidemiological studies)
- Here is where the temptation for selective reporting is maximal! Avoid at all cost.

Impact

- Present information on population attributable fraction due to genetic variant(s) of interest
- In the interpretation, consider that obtaining genetic information is a form of screening and “by definition all screening does some harm, some screening does also some good”